Plan Online, Learn Offline: Efficient Learning and Exploration via Model-Based Control

Kendall Lowrey*, Aravind Rajeswaran*,
Sham Kakade, Emanuel Todorov, Igor Mordatch

PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING
OpenAI

Problem Setting

Agent dropped into complex world, knows nominal dynamics.

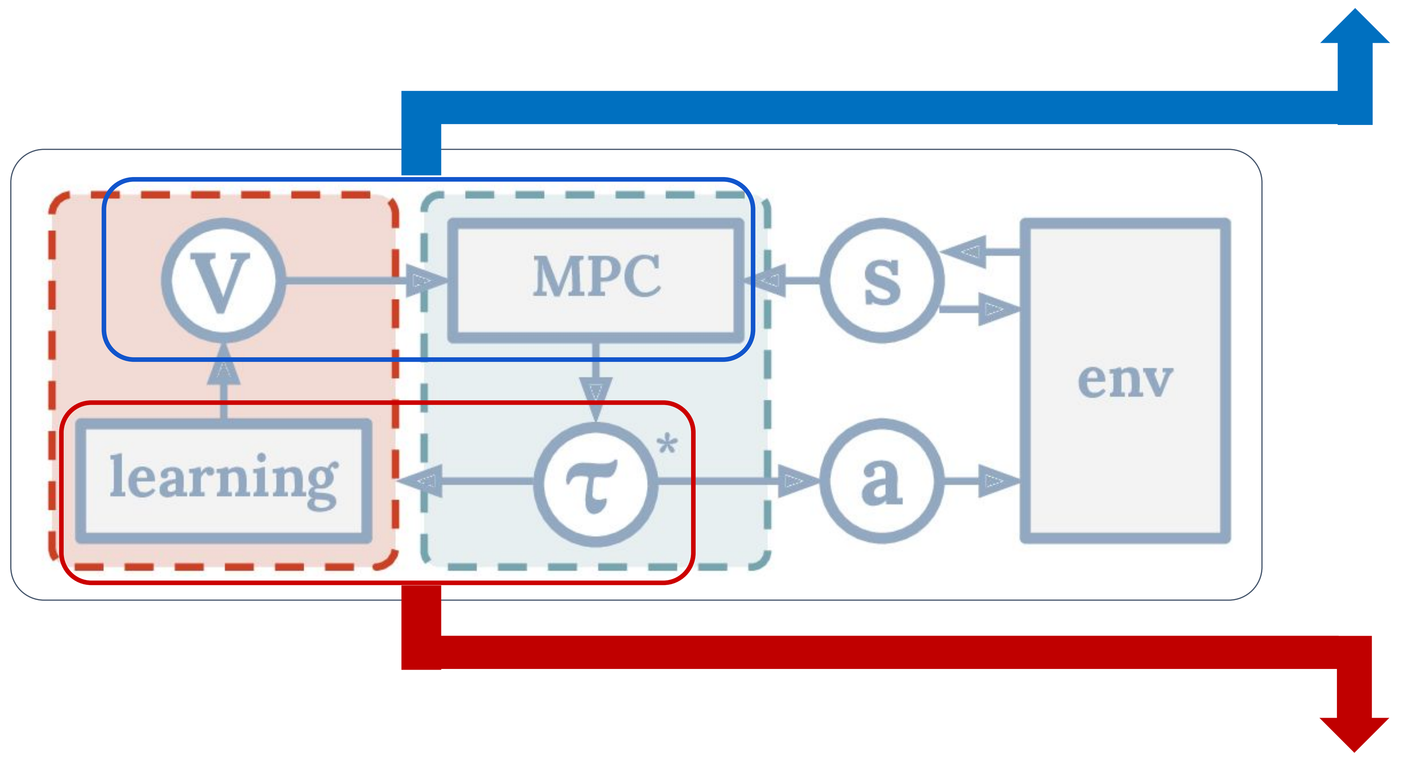
Components of an efficient learning algorithm?

- **Online Optimization** for fast and efficient improvements
- **Consolidation** of experience to enable faster and longer term planning
- **Directed Exploration** to efficiently discover optimal behaviors

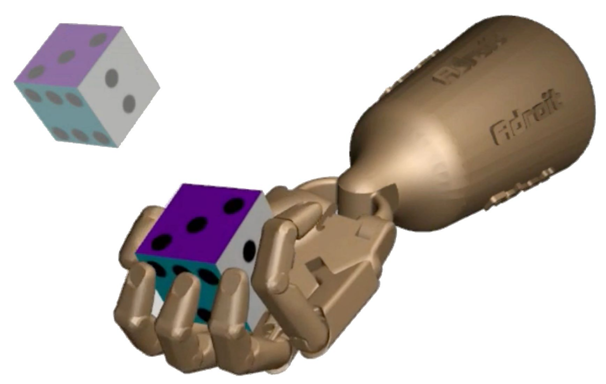
Components of POLO

Model predictive control + Terminal value function:

$$\hat{\pi}_{\text{POLO}}(s) := \underset{a_{0:H}|s_0=s}{\text{argmax}}\ \mathbb{E}\left[\underbrace{\sum_{t=0}^{H-1}\gamma^t R(s_t,a_t)}_{\text{(short horizon bias) MPC}} + \underbrace{\gamma^H \hat{V}(s_H)}_{\text{Terminal value}}\right]$$



Model-based multi-step improvement procedure



Multi-step value backup through MPC provides stable targets for FVI with minimal drift, and helps convergence

Fitted value iteration:

$$\theta_{k+1} = \underset{\theta}{\text{argmin}}\ \mathbb{E}_{s\sim\nu}\left[\left(V_\theta(s) - y(s)\right)^2\right] \quad \text{where}$$

$$y(s) := \max_{a_{0:N}|s_0=s} \mathbb{E}\left[\sum_{t=0}^{N-1}\gamma^t R(s_t,a_t) + \gamma^H V_{\theta_k}(s_N)\right]$$

Planned exploration with optimism in face of value uncertainty:
Train multiple (ensemble) value networks and form optimistic estimate of the value function. Use this optimistic estimate as terminal value in MPC

$$\hat{V}(s) := \sum_{i=1}^{M}\omega_i(s)V_{\theta_i}(s), \ \text{where}\ \omega_i(s) \stackrel{\text{def}}{:=} \frac{\exp\left(\kappa V_{\theta_i}(s)\right)}{\sum_{j=1}^{M}\exp\left(\kappa V_{\theta_j}(s)\right)}$$

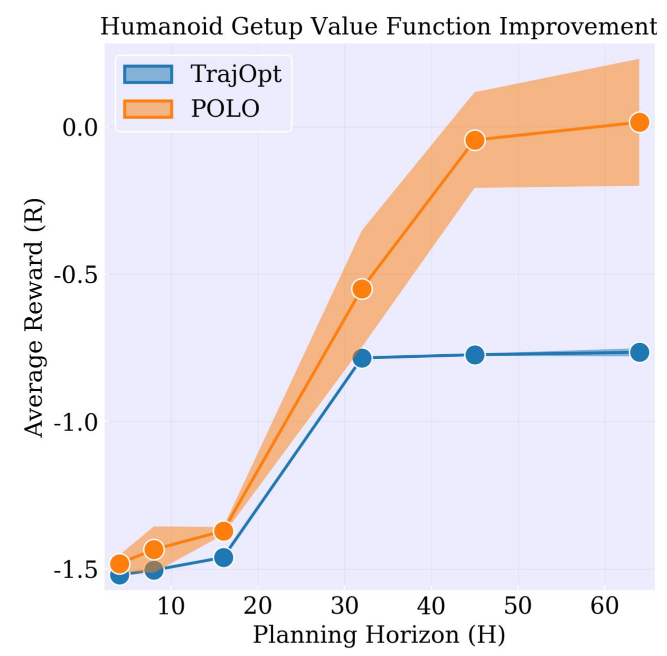
Theoretical Results

$$J^\beta(\pi) := \mathbb{E}_{s\sim\beta}\left[V^\pi(s)\right] \qquad \Delta^\beta := J^\beta(\pi^*) - J^\beta(\pi) \qquad \epsilon := \max_s |\hat{V}(s) - V^*(s)|$$

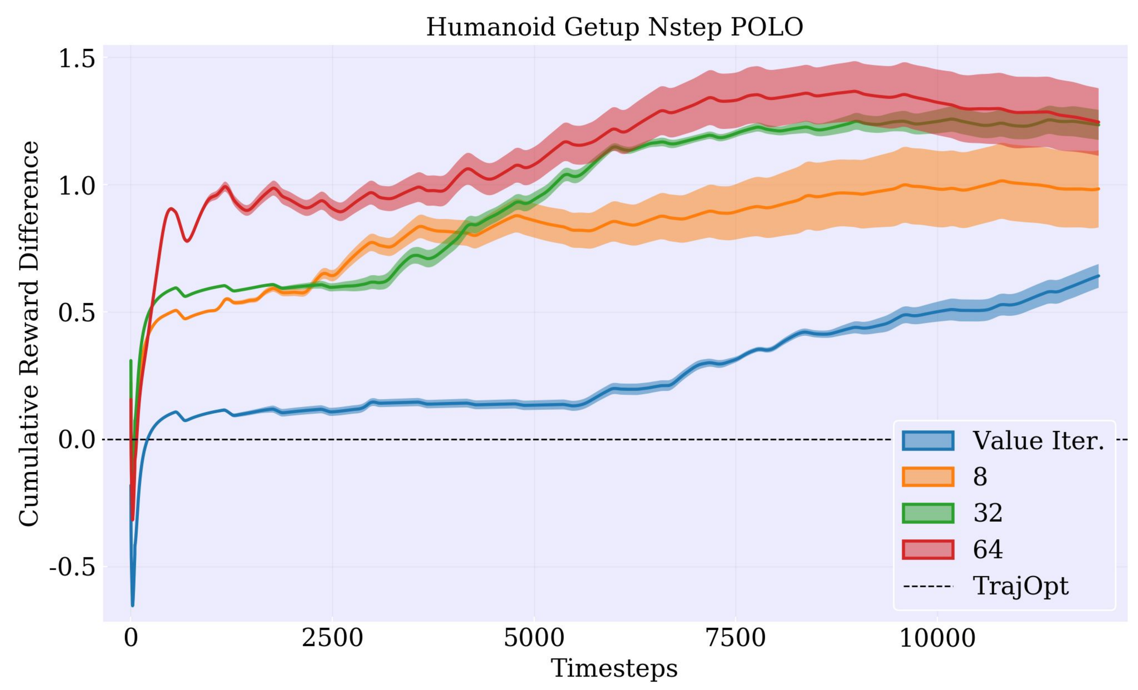
Greedy policy with FVI	Pure MPC (no VF)	POLO
$\Delta^\beta = \theta\left(\frac{\gamma\epsilon}{1-\gamma}\right)$	$\Delta^\beta \leq \frac{2\gamma^H r_{max}}{(1-\gamma)(1-\gamma^H)}$	$\Delta^\beta \leq \frac{2\gamma^H\epsilon}{1-\gamma^H}$

For $H > 1$ and non-trivial ϵ, POLO strictly better than FVI and pure MPC

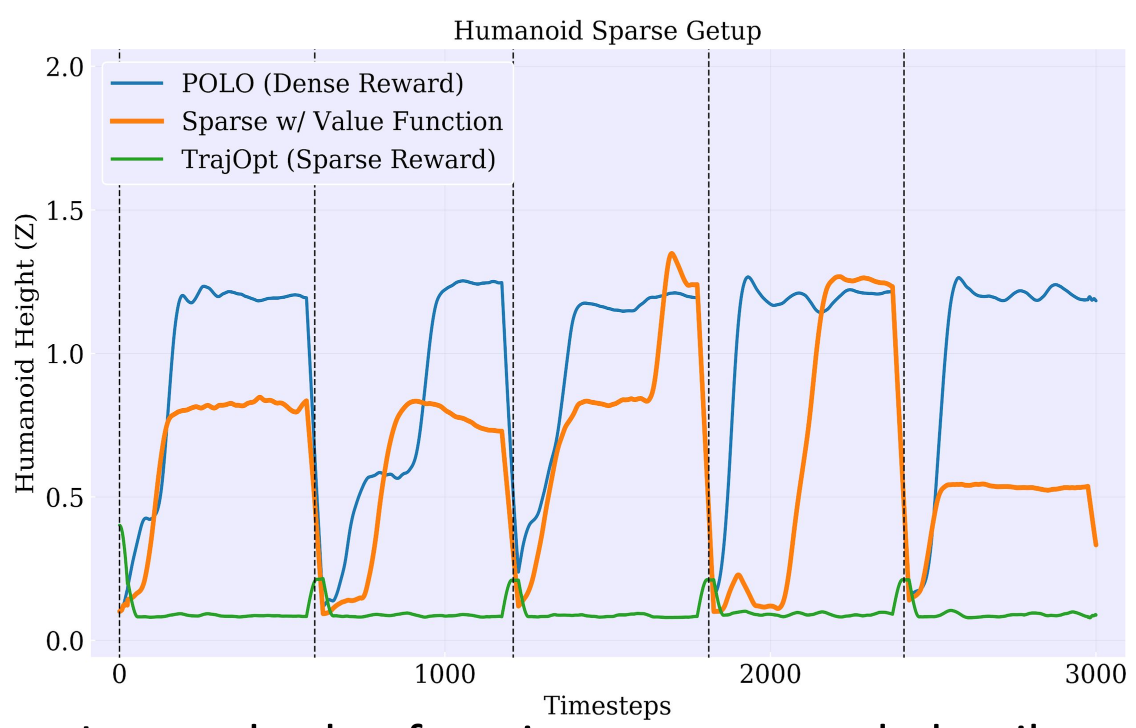
Empirical Results and Analysis



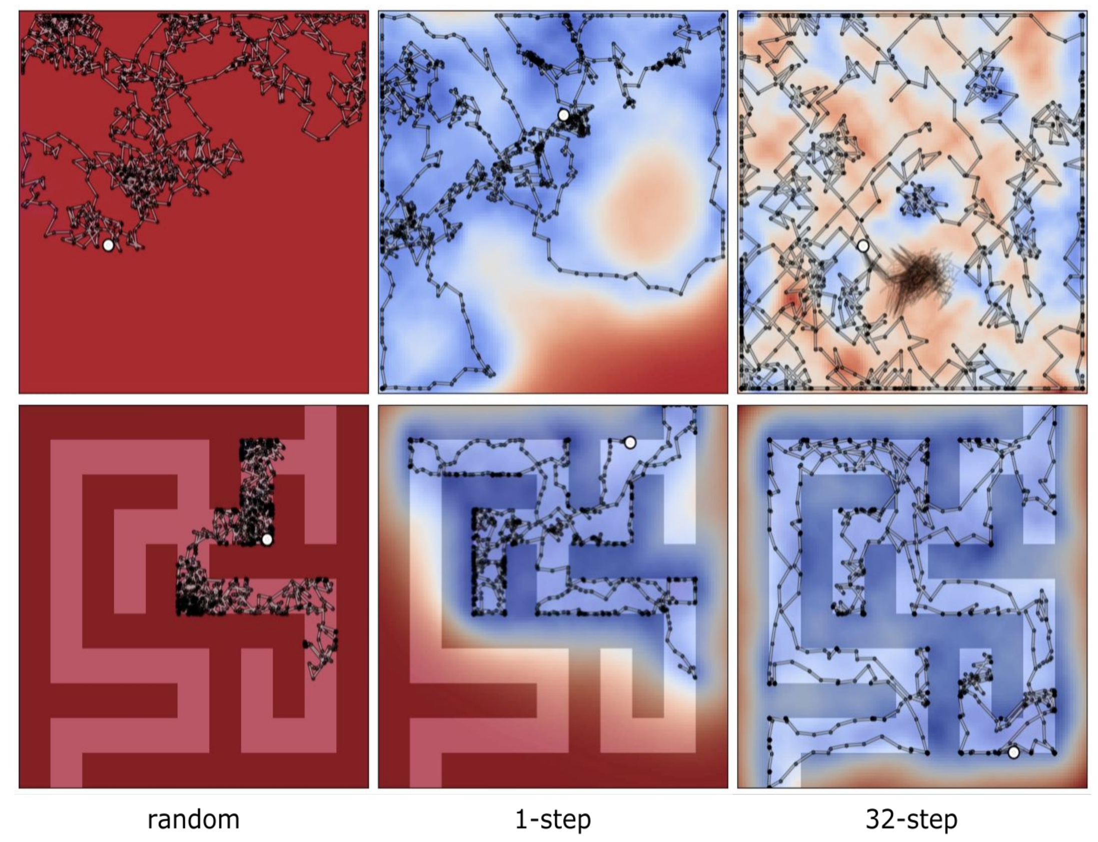
POLO finds better, non-local solutions



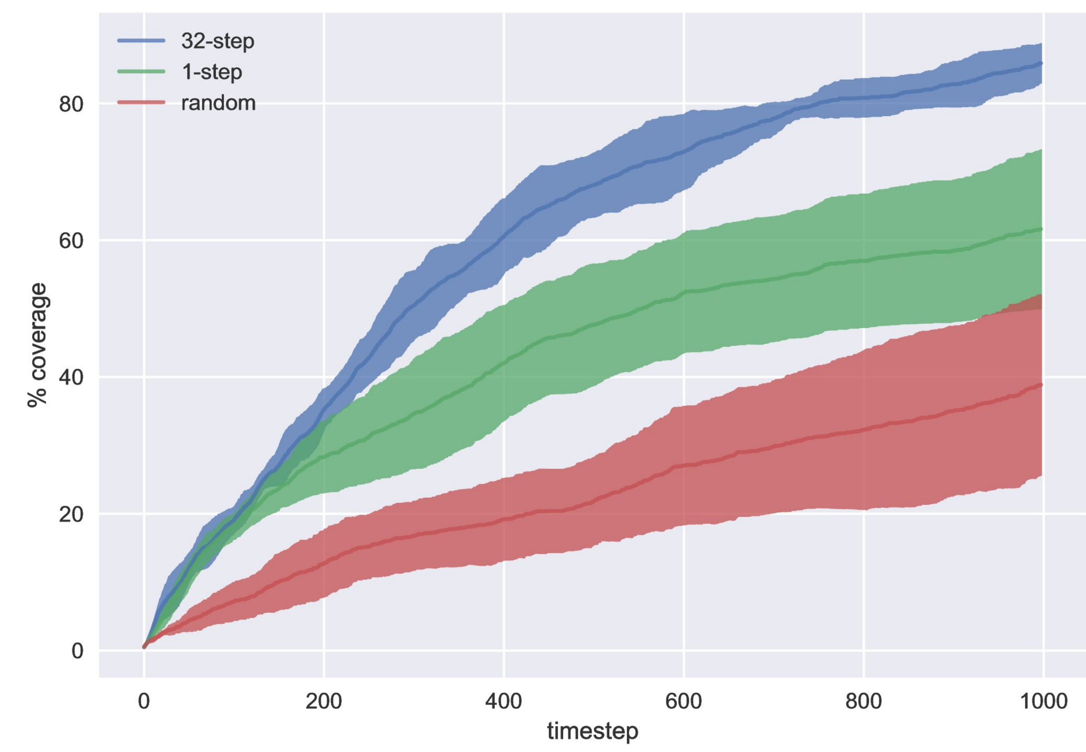
N-step targets (from MPC) stabilize value learning



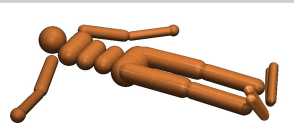
Learned value function captures task details



Planned exploration over value function uncertainty enables faster coverage



POLO uses online **optimization** for fast and efficient adaptation, **consolidates** collected experience into learned value function, and employs directed **exploration** to efficiently discover global solutions..



(* equal contributions) {klowrey, aravraj} @ cs.uw.edu | University of Washington, OpenAI